DOCUMENT RESUME

ED 245 014 UD 023 595

AUTHOR Gonzalez-Tamayo, Eulogio
TITLE Functioning-Level and Grade-Level Testing in Compensatory Educational Programs.
PUB DATE 83
NOTE 12p.
PUB TYPE Reports - General (140)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Elementary Education; *Evaluation Methods; *Grade Equivalent Scores; *Program Evaluation; *Test Validity
IDENTIFIERS *Functioning Level Testing

ABSTRACT

Functioning-level testing has been proposed as an alternative to grade-level testing for the evaluation of compensatory educational programs in which students score well below their own grade level. Data from two studies, however, suggest an adverse impact on program evaluation from the adoption of functioning-level testing. Two compensatory programs in New York City public schools were evaluated in 1974, and both functioning- and grade-level data were collected. In the first (involving fifth grade students), data confirmed both a lower number of guessing scores and lower group mean, when the functioning-level results were used. This difference stems not from a difference in guessing, however, but from the fact that the functioning-level tests do not necessarily measure subject matter taught in the compensatory program. In the second study, second graders tested at their grade level demonstrated greater pre- and post-program improvements than students in the same program and same district who were tested at functioning-level. Clearly, the difference was due to the tests themselves. Finally, in all schools evaluated, individual raw scores were routinely transformed into Grade Equivalent (GE) scores, and the individual GE scores were averaged to find the mean for the school. This practice is technically wrong and can create artificial evaluation results. (GC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

4/25/84

ED245014

FUNCTIONING-LEVEL AND GRADE-LEVEL TESTING IN COMPENSATORY EDUCATIONAL PROGRAMS

Eulogio Gonzalez-Tamayo

New York, 1983

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Eulogio Gonzalez Tamayo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

UD023595

SUMMARY

Functioning-level testing as opposed to grade-level testing has been proposed to evaluate compensatory educational programs where students score well below their own grade level. Data are presented suggesting the adverse impact on program's evaluation from the adoption of functioning-level testing. The practice of finding the group mean from individual Grade Equivalent scores is rejected as technically unsound.

Compensatory educational programs were designed to improve the educational skills, especially reading and math, of those students scoring well below their own grade level. Evaluating these programs presented both an educational and a psychometric problem. If the students were tested with tests developed for their own grade level, it could create frustration because the tasks included in the test were above their achievement level and it would also produce questionable test scores because of the large amount of guessing involved when students do not know most of the answers. If the students were tested at their own functioning level, it was thought, both problems would be solved and the functioning-level testing was, therefore, instituted in some schools.

Wick (1983) reported that changing from grade to functioning-level testing, a) lowered the average test scores and b) substantially reduced the percentage of chance level scores for students in compensatory programs. He also implied that functioning-level scores are more valid than grade-level scores.

In 1974 we had the opportunity to evaluate some of these compensatory programs in the New York City public schools. We collected scores both at the functioning and the grade level for the same or comparable groups of students during the same school year. Wick's (1983) data, on the other hand, are for different students and two different school years.

First Example

In one of the schools, fifth grade students were pre-tested with the Elementary Level of the Metropolitan Achievement Test (MAT, 1971), one level below grade-level (functioning level), and

post-tested with the Intermediate Level or grade-level. Two different levels during the school year was not the ideal way to measure actual improvements; accordingly, the students were given the Elementary level, the same level of the pre-test, three days after being tested with the Intermediate level. The data reported in Table 1 are for the Intermediate and the Elementary levels used with only three days interval.

There was a smaller amount of guessing with functioning-level testing (67% and 57% average number of correct answers for word knowledge and reading respectively) than with grade-level testing (39% and 41% average number of correct answers.) The means at the functioning and grade levels, however, were in the fat part of the normal distribution which discriminates better around the mean level of difficulty. The data confirmed both Wick's(1983) findings: lower number of guessing scores and lower group mean using the functioning than the grade level tests. We disagree, however, with the conclusion that functioning-level testing is more valid than grade-level testing. Validity is more than reduction of errors (an errorless test could be a useless test); most importantly, it implies that the test is measuring what it is supposed or intended to measure.

The difference in grade equivalent mean scores between the functioning and the grade-level tests for the same group of students with three days interval between tests can not be explained by guessing alone. First, the difference is too large, about four months in word knowledge and one year in reading; students scored higher at their own grade-level than at the functioning-level. Second, because of the regression of the means, it was expected that the students would score higher on the second than on the first testing; the result was the opposite, with the mean score being higher the first (their

own grade-level) than the second time three days later. Third, we do not see any reason to believe that the mean score would be lower if chance level score were eliminated (19% of the scores in the present case). On the contrary, keeping only the scores above chance level would raise the group mean; changing the level because some of the students scored at or below the chance level would penalize those students with scores above the chance level.

There is always some guessing in any test depending on the difficulty level of the items. In the national sample for the MAT, there is a certain percentage of students scoring at the chance level, 10% and 11% in word knowledge and reading respectively in the Elementary Level (end of third grade) and 11% and 18% in the Intermediate Level (beginning of fifth grade).

If the elimination of chance scores and errors of measurement were our only objective in a test, we could easily achieve it by testing students with lower levels; if this is done, guessing will be reduced and the mean group score will be lower. A national sample of ninth grade students taking the Elementary Level of the MAT have to miss, as an average, only one item in word knowledge and not more than two in reading in order to have a mean score at the same level than their mean score if they have taken the Intermediate Level. This is very unlikely and their mean score would be lower with the Elementary than with the Intermediate level.

The change from grade to functioning-level testing is predicated on the assumption that tests used in the school do not measure subject areas, as if the tests were ability instead of achievement tests. Ability would be a skill marginally related to content while achievement would be performance on a particular subject area taught in class. Most tests used in evaluating com-

pensatory programs are achievement tests. As reported in the Manual (MAT, 1971), the content of the tests is based on the subject areas taught in the classroom and it was developed after a thorough survey and study of the text-books most widely used in American schools. Since this is the case, it does not make sense to test the students with a lower level than their actual grade in the school unless the text-books are also those of the lower level. Any group will score lower when tested on a subject area not taught in the school than when they are tested in an area where they have received instruction.

Second Example

In another district, all second graders were pre-tested with the Primary Level II, the proper level for their grade, but some schools post-tested the students with the same level and other used the Primary Level I (one level below their grade).

As reported in table 2, the improvements of students tested at their grade level were much larger than the improvements made by those tested at their functioning-level. The students tested at their own grade level improved nine months in word knowledge and six in reading. Those tested at the functioning-level did not show any substantial improvement (one month during four and one half months of school); the gap between their reading level and the national norms was wider at the end than at the beginning of the program. This difference between schools in the same district, in the same program, after only four months of school, and for students in second grade, is very unlikely the result of actual differences in reading and improvement and, more likely, they reflect, as in the previous case, the effects of the tests used in the school.

The Use of GE Scores

In all schools that we evaluated we noted, as mandated by the State, the practice of transforming the individual raw scores into Grade Equivalent scores and the individual GE scores were averaged in order to find the mean for the school. The practice is technically wrong and it could create artificial results in the evaluation.

The GE scores were developed for the group and they have no meaning or value when applied to individuals. The MAT Manual is clear in this respect; "the most valid use of the GE scores is to interpret the average score made by some group of students"(p.12);"in obtaining the GE scores, the raw scores were equated before conveting them to Grade Equivalent; then and only then, the median for each grade was found:(13);"it is strongly urged that the use of Grade Equivalents be restricted to the interpretation of averages"(p.14).

The practice could produce misleading results. The scores discriminate fairly well around the mean and very poorly in the extremes; students with very low or very high scores can obtain the same GE score eveb if their raw scores are different. In our evaluations, in a group of low scorers, we found that the mean was four months lower (N = 163) when it was calculated from raw scores and transformed into GE scores (3.8 GE) than when the mean was obtained from the individual GE scores (4.2 GE). The difference between both methods is usually as great or greater than the improvements made during the program and is due exclusively to a statistical manipulation. In an hypothetical case of high scorers (we did not find this group in our evaluations), as reported in table 3, the students as a group have a mean score higher (four months higher) doing the calculations from the mean raw scores than from the average GE scores.

Table 1: Mean Scores (GE) for Fifth Grade Students (N=26) Tested at Grade-level and at Functioning-level

Test	Functioning Level	Grade Level	Difference
Word Knowledge	3.85 (67%)	4.20 (39%)	0.35
Reading	3.80 (57%)	4.80 (41%)	1.00
National level	5.80	5.80	

N.B. The percentages in parenthesis indicate the average number of correct answers for the means.

Table 2: Mean Scores (GE) for Second Grade Students Testes at Grade-Level (N=197) and at Functioning-Level (N=644)

Test	Grade-Level		Functioning-Level		Mean Improvement	
	Pre-test	Post-test	Pre-test	Post-test	Grade	Functioning
Word Knowledge	1.98	2.89	1.79	1.90	0.91	0.11
Reading	1.92	2.54	1.79	1.93	0.62	0.14
National Level	2.40	2.85	2.40	2.85	0.45	0.45

Table 3: Mean Total Reading from the MAT for High Scorers (Elementary Level) (Hypothetical case).

N	Raw Score	GE Score
10	86	7.2
10	87	7.6
10	91	9.8
10	94	9.8
Average	89.5= 9.0 GE	8.6 GE

BIBLIOGRAPHY

1. Durost, W. N., Bixler, H. H., Wrightstone, J. W., Prescott, G. A., Balow, I. H. Metropolitan Achievement Tests (MAT). New York, Harcourt Brace Jovanovich, 1971.

2. Wick, J. W. Reducing Proportion of Chance Scores in Inner-City Standardized Testing Results: Impact on Average Scores. American Educational Research Journal, 1983, 20, 461-463.